ACTO DE INVESTIDURA DEL GRADO DE DOCTOR HONORIS CAUSA

VINTON G. CERF
RICHARD SCHROCK

STVDIVM GENERALE CAESARAVGVSTANAE CIVITATIS

Prensas Universitarias de Zaragoza

ALOCUCIÓN DEL PROFESOR DR. VINTON G. CERF

Rector Magnífico
Academic Authorities
Distinguished Doctors
Ladies and gentlemen

There can be little doubt that as the 21st century unfolds, an ever-increasing amount of information is flowing into the Internet's World Wide Web or into digital media such as CD-ROMs, DVDs and other digital recording media. Moreover, these digital representations of content are increasingly complex objects, not necessarily easily rendered in simple ways by printing or even electronic display.

Some of these digital objects are multi-media in scope (sound, imagery, motion). Some of them incorporate programmable elements such as spreadsheets, complex documents, electronic mail containing attachments and embedded complex objects. Some of the digital content that is found on the Internet (or in private networking

systems and other digital media) are interpretable only by special purpose application software. Some of the objects are in fact software in and of themselves. Think of Java programs downloaded in the course of surfing the Internet with a browser. These software snippets are meaningless unless interpreted in the proper context and with the proper software. Other objects are just text until they have been processed by compilers and loaded into a computer for interpretation.

The implications of this increasingly diverse digital universe of content are profound. In the absence of appropriate software and hardware platforms, these digital bits are in fact useless data. I have taken to referring to the potential loss of interpretability as "bit rot" and I worry that as the decades unfold, the Internet and all other digital media that hold such data will slowly disintegrate into a massive pile of useless, uninterpretable, meaningless data: a pile of rotten bits.

This is not just a problem for the historians and archeologists of the far future although I do harbor nightmares that our descendants will wonder what on earth the 20th and 21st centuries were about and will be unable to find out because we have left behind a massive midden of bit rot. This is already a current problem for many users of digital technology. Have you ever upgraded your computer operating system or perhaps bought a new computer only to discover that some of the digital photographs you could view with older software sudden are impossible to display because of software incompatibilities or lack of backwards compatibility? Have you ever tried to display a presentation produced with an earlier version of software using the latest version, only to discover that the older version is not recognized or is misinterpreted? Can you imagine what might

happen if, in the year 3000 you were searching the then "Internet" and found a data file from the 21st century but had no way to display it because the software that produced it was long lost? These are all real problems today and will grow increasingly serious as time continues its inexorable march into the future.

What is of particular concern is that these complex objects yield their full value only when the appropriate software is available to permit their manipulation, editing, presentation and use. We no longer have the luxury of preserving these objects by printing or audio or video recordings because the full extent of their interpretation requires much more elaborate contextual treatment using the appropriate application software.

These concerns do not even take into account the more obvious problem of physical formats of material falling into disuse. If you can recall the use of 8 or 5 1/4 "floppy disks" or the 3 1/2 successors, you might wonder whether there is any equipment currently available that can read these formats. If there is any left, it may well be in a museum and non-operational. The CD-ROMs of today will give way to the DVDs and HD-DVDs of tomorrow and data recorded on the older formats will become inaccessible. The physical media themselves may deteriorate as well. Who knows how long a polycarbonate CD will actually last?

In this essay, I will set aside the problem of physical deterioration on the grounds that one can reasonably imagine copying bits from an older digital medium to another. We see this today as movies are re-copied from film to digital DVDs or digital tape, for example. I want to concentrate on the problem of digital data formats and their interpretation over long periods of time, measured in decades to centuries.

Software evolves as use and invention dictate. New functionality often demands new and expanded formats. Some changes may not be backward compatible, rendering earlier data files uninterpretable. Even when this is not the case, it seems predictable that some software will fall into disuse and will no longer be maintained. The company or persons supporting the software may cease doing so for any of a variety of reasons.

The implications of these observations are both obvious and subtle. First, it seems important that software that ceases to be commercially available or supported ought to be placed into a category that permits users access to it in some fashion, perhaps through online "cloud computing" services. Moreover, these applications may require certain versions of operating systems to operate properly so the same concerns about application software may apply to operating systems. One could even go so far as to argue that the hardware on which the operating system and application software runs needs to be preserved or emulated so that their functionality is not lost in the future.

There are obvious intellectual property issues associated with these questions. Will the source code be made available when a particular version of application or operating software is not longer supported? If not, is there a way to make the software accessible on the Internet, for example, so that their functionality is preserved? While there is likely no panacea to assure that all digital information can be preserved over time, it strikes me as important to take steps now to develop techniques and global policy regimes that support this goal.

Much has been made, thus far, of the format of digital data and its interpretability by software. Perhaps it is worth turning to another aspect of this challenge having to do with the semantics of information. The meaning of

words, symbols and digital formats are central to their utility. Anyone who has sought to read written material from the relatively distant past can appreciate how language evolves and its meanings change. Even in contemporary terms, there is ambiguity that can only be resolved with proper context. In English, the term "red" can mean a color or in some contexts a political persuasion, for example. It is reasonable to ask whether there is any way to establish a general digital semantics that would allow long-term interpretation of digital content even with the evolution of the application software that interprets it.

There have been attempts in the past to create long-lasting representations of the syntax of digital information. One example of this is the so-called Abstract Syntax Notation (Version 1) or ASN.1[1].

This relatively complex standard is intended to allow very general descriptions of complex digital objects so that these descriptions are in some sense self-defining. Without debating the general utility of these methods, it is fair to observe that even when they convey formatting information accurately, there remains the problem of interpreting the meaning of content represented in these formats. A semantic complement to the syntax representation would be an interesting technical challenge. It is almost certain that some efforts have been made in the past or are being made in the future.

Another example of an attempt to capture syntax and meaning is found in the Extensible Markup Language (XML)[2] that has as its roots both the Hypertext Markup

1 <http://asn1.elibel.tm.fr/>.
2 <http://asn1.elibel.tm.fr/>.

Language (HTML)[3] of the World Wide Web and the earlier Standardized Generalized Markup Language (SGML)[4] from which HTML is derived. Another variation on this theme is the invention of programming languages whose compilers can be programmed in the same language. SmallTalk[5] is one example; SQUEAK[6], PYTHON[7] and JAVA[8] are others.

It is tempting to speculate whether these high level programming languages could be used in conjunction with various syntax defining standards to capture the semantics of syntactic structures in ways that would preserve both the syntax and semantics of digital objects. It seems likely that work is on-going in this area of research and I would urge those interested in the problem to carry out further exploration in the worlds comprising computer science, digital library design, and computer languages.

There are additional deficiencies in our present world of digital objects and one of the most pressing to remedy is the way in which we make reference to online information. As should be obvious from the footnotes to this paper, references to WWW-based information is frequently by use of so-called Uniform Record Locators (URLs)[9]. In fact, in the language of the World Wide Web, the terms "Uniform Record Locator," "Uniform Record Identifier (URI)," and "Unform Record Name (URN)"

3 <http://www.w3.org/TR/REC-html40/>.
4 <http://www.w3.org/TR/html401/intro/sgmltut.html>.
5 <http://www.smalltalk.org/main/>.
6 <http://www.squeak.org/>.
7 <http://www.python.org/>.
8 <http://java.sun.com/>.
9 <http://www.w3.org/Addressing/>.

are all intended to convey different ways to identifying digital objects found on the Internet.

It is generally the case that all such references depend on some kind of "lookup" to translate the reference into a definite place on the Internet at which to look for the referenced object. Many of these objects depend on the Domain Name System (DNS)[10] that translates Domain Names such as <www.google.com> into specific, numerical Internet Addresses such as 209.85.173.103. It should be obvious that this mapping function from name to address has the potential hazard that material referenced by domain name may become inaccessible if the computer at the target destination no longer holds the data or if the domain name becomes unregistered or is re-purposed in the future. These references are therefore in some sense ephemeral and do not satisfy the desire for longevity of reference that digital librarians, historians and others will rely upon in the future.

To be fair, the concept of Unform Resource Name is intended to provide a reference that is unchanging with regard to the Domain Name System. However, this concept relies on the idea that something will be able to map the URN into a meaningful Internet address (or the equivalent) far into the future.

One attempt at this is called the Handle System[11] developed by the Corporation for National Research Initiatives. In this system, generalized numeric identifiers are mapped through a distributed, replicated directory system into references to a replicated, distributed system of servers. The digital objects registered and contained in

10 <http://www.dns.net/dnsrd/>.
11 <http://www.handle.net/>.

the Handle System have substantial meta-data associated with them including information as to the source of the object, terms and conditions for access to it, formatting information and other critical data needed to use the object. The system serves multiple purposes. It can be used as a mechanism for tracking intellectual property rights, terms and conditions; it can be used to find objects in perpetuity (at least as long at the Handle System itself is maintained and operated). In part, the motivation for the design and invention of the system was to provide object identifiers that are persistent over long periods of time and not subject to invalidation by changes in domain names and their mappings to Internet addresses.

Implicit in the Handle System design is the interesting problem of establishing long-lived and interpretable meta-data about the digital object and its characteristics. The problem of designing an extensible syntax and semantics for this meta-data is another instance of the problem of persistent meaning. The deeper one penetrates into the problem, the larger it seems to grow. For example, references to the owners of digital objects or holder of rights in them begs the question how to express these references in long-lived terms. How can we track the holder of rights over periods of decades? How can we reference the sources of these objects over centuries? These and other questions form a tapestry of difficult, important and useful research questions.

Placing these questions into an international context that includes all the languages of the world adds the question of language representations in digital form. The Unicode[12] table of scripts for the world's languages is itself a major undertaking. Its coding of characters can form a common framework for the representation of information in all the world's languages. The HTML and XML mentioned above make use of this important and evolving table. Language itself changes and evolves as do alphabets over centuries. Looking at this problem from the perspective of a thousand years or more produces a sobering view of the magnitude of the problem of stable referencing in the digital world.

12 <http://unicode.org/>.

One is tempted to suggest that these questions are part of a larger notion of information ecology in which institutions, technology, society and the global economy play important and dynamic roles. That all of these moving parts need to be coordinated through some underlying organizing principles represents one of the major challenges of the Internet and its global utility. There is ample room for a great deal of experimentation, scholarly research and organizational collaboration.

I hope that these brief remarks will trigger interest in the minds of scholars in search of serious dissertation topics or those dedicating themselves to the longevity of the online universe. Success in this work will benefit generations to come and offer to our descendants the opportunity to appreciate and even experience the digital world of this century. It is our way to communicate with the populations of the future and to convey to them our hopes, fears, beliefs, successes and mistakes. While we cannot ourselves peer deeply into the future, we can at least offer the future an opportunity to see deeply into their digital past with a clarity that I hope will be appreciated and perhaps also essential to their understanding of their own digital world that we can only dimly imagine.

Vinton G. CERF